第27卷 第5期
2010年10月

工 程 数 学 学 报
CHINESE JOURNAL OF ENGINEERING MATHEMATICS

Vol. 27 No. 5
Oct. 2010

文章编号:1005-3085(2010)05-0883-06

集值信息系统的属性约简*

马建敏[1]， 朱朝晖[2]

(1- 长安大学理学院数学与信息科学系，西安 710064； 2- 深圳卓成混凝土模块研究所，深圳 518000)

摘 要: 属性约简是粗糙集理论研究中的重要内容之一。本文主要研究集值信息系统的属性约简问题。在集值信息系统中基于拟序关系引入了信息量的概念，给出了属性特征的判定方法，以及信息量与属性约简之间的关系。根据信息量定义了属性重要性，研究了属性重要性与属性约简之间的关系。进而得到了基于信息量和属性重要性的属性约简算法，给出了该算法的时间复杂度。通过实例说明，该算法是有效的。

关键词: 集值信息系统；拟序关系；信息量；属性重要性；属性约简

分类号: **AMS**(2000) 65L07; 65N12 **中图分类号**: TP18 **文献标识码**: A

1 引言

粗糙集理论是由波兰数学家Pawlak于1982提出的一种数据分析理论[1]。该理论由于能分析处理不精确、不协调和不完备等信息引起人工智能工作者的广泛关注，并被成功应用在机器学习与知识发现、数据挖掘、决策支持与分析、过程控制、模式识别等领域[2]。

属性约简作为粗糙集理论的重要研究内容之一[1,3]，是在保持分类能力不变的前提下删除其中的冗余属性。由于属性约简并不唯一，人们希望找出所有约简或最小约简。但寻找最小约简是NP-hard问题[4]。解决这类问题的一般方法是采用启发式搜索方法求出最优或次优约简[5]。苗夺谦等人[6]提出了基于互信息的知识相对约简的启发式算法。王国胤等人提出了基于条件信息熵的决策表约简算法[7]。梁吉业等人[8]提出了基于信息量的属性约简算法。黄兵等人[9]给出了不完备信息系统的属性约简算法。而对不确定或缺省信息，则需研究不完备信息系统或集值信息系统。

本文在集值信息系统中建立了拟序关系，由此引入了信息量的概念，通过信息量研究了属性特征，以及信息量与约简之间的关系。进而给出了属性重要性的定义，研究了属性重要性与约简之间的关系。并基于信息量和属性重要性给出了获取集值信息系统的属性约简的算法。通过实例验证了该算法的有效性。

2 集值信息系统

定义1 称(U,A,F)为集值信息系统，其中$U=\{x_1,x_2,\cdots,x_n\}$是非空有限对象集合，称为论域；$A=\{a_1,a_2,\cdots,a_m\}$是非空有限属性集合；$F=\{f_a:\forall\, a\in A\}$是$U$到$A$上的函数集合，其中$f_a:U\to\mathcal{P}_0(V_a)\,(\forall\, a\in A)$称为信息函数，$V_a$是属性$a$的值域，$\mathcal{P}_0(V_a)$是$V_a$上非空子集的全体。

收稿日期: 2009-07-20. **作者简介**: 马建敏(1978年3月生)，女，博士，讲师. 研究方向：概念格与粒计算.
***基金项目**: 国家自然科学基金(10901025)；中央高校基本科研业务费专项基金(CHD2009JC028).

定义2 设 (U,A,F) 为集值信息系统。对任意的 $B\subseteq A$，定义 U 上的二元关系

$$R_B^{\subseteq}=\big\{(x,y)\in U\times U: f_a(x)\subseteq f_a(y), \forall\, a\in B\big\},$$

称 $R_B^{\subseteq}$ 为集值信息系统 (U,A,F) 上的拟序关系。

显然，$R_B^{\subseteq}$ 是自反、传递的，但不是对称的，故不再是等价关系。记

$$R_B^{\subseteq}(x)=\big\{y\in U:(x,y)\in R_B^{\subseteq}\big\},$$

称 $R_B^{\subseteq}(x)$ 为包含 x 的信息粒，则全体信息粒 $U/R_B^{\subseteq}=\{R_B^{\subseteq}(x): x\in U\}$ 构成 U 的一个覆盖。

性质1 设 (U,A,F) 为集值信息系统，$R_B^{\subseteq}$ 为 (U,A,F) 上的拟序关系，则对任意的 $B\subseteq A$, $x,y\in U$，有

1) 若 $B\subseteq A$，则 $R_A^{\subseteq}\subseteq R_B^{\subseteq}$ 且 $R_B^{\subseteq}=\bigcap\limits_{a\in B}R_a^{\subseteq}$；$R_A^{\subseteq}(x)\subseteq R_B^{\subseteq}(x)$ 且 $R_B^{\subseteq}(x)=\bigcap\limits_{a\in B}R_a^{\subseteq}(x)$；

2) 若 $y\in R_B^{\subseteq}(x)$，则 $R_B^{\subseteq}(y)\subseteq R_B^{\subseteq}(x)$ 且 $R_B^{\subseteq}(x)=\cup\{R_B^{\subseteq}(y): y\in R_B^{\subseteq}(x)\}$；

3) $R_B^{\subseteq}(y)=R_B^{\subseteq}(x)\Leftrightarrow\forall\, a\in B,\ f_a(x)=f_a(y)$。

定义3 设 (U,A,F) 是集值信息系统。对任意的 $a\in A$，若 $R_{A-\{a\}}^{\subseteq}=R_A^{\subseteq}$，则称属性 a 在 A 中是不必要的；否则，称 a 在 A 中是必要的。若对每个 $a\in A$ 在 A 中都是必要的，则称 A 是独立的，否则，称 A 是相依的。

定义4 设 (U,A,F) 是集值信息系统，A 中所有必要属性组成的集合称为属性集 A 的核，记作 $\mathrm{Core}(A)$。

定义5 设 (U,A,F) 是集值信息系统，$B\subseteq A$。如果 $R_B^{\subseteq}=R_A^{\subseteq}$，则称 B 为集值信息系统 (U,A,F) 的协调集；若 B 为 (U,A,F) 的协调集，且对任意的 $a\in B$，$R_{B-\{a\}}^{\subseteq}\neq R_A^{\subseteq}$，称 B 为 (U,A,F) 的约简。

易证，$\mathrm{Core}(A)=\cap\{D: D\subseteq A,\ D$ 是 (U,A,F) 的约简$\}$。

3 集值信息系统上的信息量及属性重要性

定义6 设 (U,A,F) 是集值信息系统，$B\subseteq A$，且 $U/R_B^{\subseteq}=\big\{R_B^{\subseteq}(x_i): x_i\in U\big\}$，则 B 的信息量定义为

$$I(B)=1-\frac{1}{|U|^2}\sum_{i=1}^{|U|}\big|R_B^{\subseteq}(x_i)\big|,$$

其中 $|X|$ 表示集合 X 的基数。

性质2 设 (U,A,F) 是集值信息系统，则对任意的 $B\subseteq A$, $I(B)\leq I(A)$。

定理1 (协调集的判定定理) 设 (U,A,F) 是集值信息系统，则对任意的 $B\subseteq A$，B 是 (U,A,F) 的协调集 $\Leftrightarrow I(A)=I(B)$。

定理2 (属性特征的判定方法) 设 (U,A,F) 是集值信息系统，则对任意的 $a\in A$，a 是必要的 $\Leftrightarrow I(A-\{a\})<I(A)$。

证明 设 a 是必要的，则 $R_{A-\{a\}}^{\subseteq}\neq R_A^{\subseteq}$。由性质1知 $R_A^{\subseteq}\subset R_{A-\{a\}}^{\subseteq}$。由定义6即得 $I(A-\{a\})<I(A)$。若 $I(A-\{a\})<I(A)$，由定义6及性质1知，存在 $x_i\in U$ 使

$$R_{A-\{a\}}^{\subseteq}(x_i)\neq R_A^{\subseteq}(x_i).$$

即 a 为必要属性。

性质 3　设 (U,A,F) 是集值信息系统，则对任意的 $B\subseteq A$，B 是独立的 $\Leftrightarrow \forall\, a\in B$, $I(B-\{a\})<I(B)$。

定理 3　设 (U,A,F) 是集值信息系统，则 $\mathrm{Core}(A)=\{a\in A:\ I_{A-\{a\}}(a)<I(A)\}$。

证明　$a\in \mathrm{Core}(A)\Leftrightarrow A-\{a\}$ 是不协调的 $\Leftrightarrow R^{\subseteq}_{A-\{a\}}\neq R^{\subseteq}_{A}\Leftrightarrow R^{\subseteq}_{A}\subset R^{\subseteq}_{A-\{a\}}\Leftrightarrow I(A)>I(A-\{a\})$。

定理 4　设 (U,A,F) 是集值信息系统，则对任意的 $B\subseteq A$，B 是约简 $\Leftrightarrow I(B)=I(A)$，且对任意的 $a\in B$, $I(B-\{a\})<I(A)$。

定义 7　设 (U,A,F) 是集值信息系统，$a\in A$。属性 a 在 A 中的重要性定义为

$$\mathrm{Sig}_{A-\{a\}}(a)=I(A)-I\big(A-\{a\}\big).$$

特别地，当 $A=\{a\}$，用 $\mathrm{Sig}(a)$ 表示 $\mathrm{Sig}_{\emptyset}(a)$，则 $\mathrm{Sig}(a)=I(\{a\})$，其中 $U/R^{\subseteq}_{\emptyset}=U$, $I(\emptyset)=0$。

性质 4　设 (U,A,F) 是集值信息系统，则对任意的 $a\in A$，有

$$0\leq \mathrm{Sig}_{A-\{a\}}(a)\leq 1-\frac{1}{|U|}.$$

定理 5　设 (U,A,F) 是集值信息系统，则对任意的 $a\in A$，

$$a\ \text{在}\ A\ \text{中是必要的}\ \Leftrightarrow \mathrm{Sig}_{A-\{a\}}(a)>0.$$

定理 6　设 (U,A,F) 是集值信息系统，则对任意的 $B\subseteq A$，B 是约简 $\Leftrightarrow I(B)=I(A)$，且对任意的 $a\in B$, $\mathrm{Sig}_{B-\{a\}}(a)>0$。

定理 7　设 (U,A,F) 是集值信息系统，则

$$\mathrm{Core}(A)=\big\{a\in A:\mathrm{Sig}_{A-\{a\}}(a)>0\big\}.$$

4　基于信息量的集值信息系统的属性约简算法

定义 8　设 (U,A,F) 是集值信息系统，$B\subseteq A$。对任意属性 $a\in A-B$，a 关于属性集 B 的重要性定义为

$$\mathrm{Sig}_B(a)=\mathrm{Sig}_{B\cup\{a\}-\{a\}}(a)=I\big(B\cup\{a\}\big)-I(B).$$

下面给出基于属性重要性的集值信息系统的属性约简算法：

输入:　集值信息系统 (U,A,F)。

输出:　集值信息系统的核与约简。

步骤 1　计算集值信息系统中知识 A 的信息量 $I(A)$；

步骤 2　$\mathrm{Core}(A):=\emptyset$。计算每个属性 a 在 A 中的重要性 $\mathrm{Sig}_{A-\{a\}}(a)$。'若 $\mathrm{Sig}_{A-\{a\}}(a)>0$，则 $\mathrm{Core}(A):=\mathrm{Core}(A)\cup\{a\}$。最后得到的 $\mathrm{Core}(A)$ 为属性集 A 的核；

步骤 3　计算核 $\mathrm{Core}(A)$ 的信息量。若 $I(\mathrm{Core}(A))=I(A)$，则输出核 $\mathrm{Core}(A)$ 即为集值信息系统的属性约简 (此时 $\mathrm{Core}(A)$ 为 (U,A,F) 最小约简)；否则，$(I(\mathrm{Core}(A))<I(A)$，执行步骤 4；

步骤 4　令 $C=\mathrm{Core}(A)$，对属性集 $A-C$ 重复执行：

1)　对每个属性 $a\in A-C$，计算属性重要性 $\mathrm{Sig}_C(a)$；

2)　选择属性 a 使其满足

$$\mathrm{Sig}_C(a)=\max_{a'\in A-C}\mathrm{Sig}_C(a'),$$

令 $C := C \cup \{a\}$;

3) 若 $I(C) = I(A)$，输出 C (此时 C 为 (U, A, F) 的一个属性约简)；否则，转 1)。

下面分析上述算法的时间复杂性：

计算核 Core(A) 共需计算 $|A|$ 次 $\mathrm{Sig}_{A-\{a\}}(a)$。

计算属性约简需要计算 $\mathrm{Sig}_C(a)$ 的次数最多为

$$|A| + \big(|A| - 1\big) + \cdots + 1 = |A|\big(|A| + 1\big)/2 = O\big(|A|^2\big).$$

为了计算 $\mathrm{Sig}_{A-\{a\}}(a)$ (计算 $\mathrm{Sig}_C(a)$ 与计算 $\mathrm{Sig}_{A-\{a\}}(a)$ 的时间复杂性相同)，需要进行下列计算：

1) 计算 $|A|$ 个覆盖。类似文献 [10] 可知，计算每个覆盖的时间复杂性为 $O(|U|^2)$，因此计算 $|A|$ 个覆盖的时间复杂性为 $O(|A| \times |U|^2)$。

2) 为了计算 $U/R_A^{\subseteq}$ 和 $U/R_{A-\{a\}}^{\subseteq}$，需要计算 $|A| - 1$ 和 $|A|$-2 次交。计算一次交的时间复杂性为 $O(|U|^2)$。因此计算这些交的时间复杂性为

$$\big(|A| - 1 + |A| - 2\big) \times O\big(|U|^2\big) = O\big(|A| \times |U|^2\big).$$

因此，计算一次 $\mathrm{Sig}_{A-\{a\}}(a)$ 的时间复杂性为 $O(|A| \times |U|^2)$。

计算 $I(A)$ 和 $I(A - a)$ 的时间复杂性为 $O(|A| \times |U|^2)$。

故整个算法的时间复杂性为

$$\big(|A| + |A|(|A| + 1)/2\big) \times O\big(|A| \times |U|^2\big) = O\big(|A|^3 \times |U|^2\big).$$

例1 表1给出了集值信息系统 (U, A, F)，其中

$$U = \{x_1, x_2, x_3, x_4, x_5, x_6\}, \quad A = \{a_1, a_2, a_3, a_4\}.$$

表1: 集值信息系统 (U, A, F)

U	a_1	a_2	a_3	a_4
x_1	{1}	{1,2}	{1}	{1,2}
x_2	{1,2,3}	{1,2}	{1,2}	{1,2}
x_3	{1}	{1}	{1,2}	{1}
x_4	{1,2}	{1}	{1,2,3}	{1}
x_5	{1,2,3}	{1,2,3}	{1,2}	{1,2,3}
x_6	{1,2,3}	{1,2}	{1,2,3}	{1,2}

下面利用属性约简算法给出集值信息系统的属性约简：

步骤1 拟序关系 $R_B^{\subseteq} = \{(x, y) \in U \times U : f_a(x) \subseteq f_a(y)\}$，则全体信息粒为

$$R_A^{\subseteq}(x_1) = \{x_1, x_2, x_5, x_6\}, \quad R_A^{\subseteq}(x_2) = \{x_2, x_5, x_6\}, \quad R_A^{\subseteq}(x_3) = \{x_2, x_3, x_4, x_5, x_6\},$$

$$R_A^{\subseteq}(x_4) = \{x_4, x_6\}, \quad R_A^{\subseteq}(x_5) = \{x_5\}, \quad R_A^{\subseteq}(x_6) = \{x_6\}.$$

故 A 的信息量

$$I(A) = 1 - \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|R_A^{\subseteq}(x_i)\right| = \frac{5}{9}.$$

步骤2 由定义7求得

$$\mathrm{Sig}_{A-\{a_1\}}(a_1) = 0, \quad \mathrm{Sig}_{A-\{a_2\}}(a_2) = 0, \quad \mathrm{Sig}_{A-\{a_3\}}(a_3) = \frac{5}{36}, \quad \mathrm{Sig}_{A-\{a_4\}}(a_4) = 0.$$

故由定理7知，$\mathrm{Core}(A) = \{a_3\}$。

步骤3 $I(\mathrm{Core}(A)) = I(\{a_3\}) = \frac{11}{36}$，且 $I(\mathrm{Core}(A)) \neq I(A)$，执行步骤4。

步骤4 令

$$C = \mathrm{Core}(A) = \{a_3\}.$$

对 $A - C = \{a_1, a_2, a_4\}$ 计算各个属性关于属性 C 的重要性：

1) $\mathrm{Sig}_C(a_1) = \frac{5}{36}$, $\mathrm{Sig}_C(a_2) = \frac{1}{4}$, $\mathrm{Sig}_C(a_4) = \frac{1}{4}$。

2) 由于

$$\mathrm{Sig}_C(a_2) = \mathrm{Sig}_C(a_4) = \max_{a\in A-C}\mathrm{Sig}_C(a) = \frac{1}{4},$$

取 $C_1 := C \cup \{a_2\}$, $C_2 := C \cup \{a_4\}$。

3) 对任意的 $i = 1, 2$, $I(C_i) = \frac{5}{9}$，且 $I(C_i) = I(A)$。

故核 $\mathrm{Core}(A) = \{a_3\}$, $C_1 = \{a_2, a_3\}$, $C_2 = \{a_3, a_4\}$ 均为集值信息系统的约简。

5 结论

粗糙集理论是一种处理不精确和不完全知识的工具，而属性约简则是粗糙集理论研究的核心问题之一。属性约简的过程即是寻找保持分类能力不变的最小属性子集。为此人们提出了基于信息熵、信息量等的属性约简算法。本文在集值信息系统中基于拟序关系提出了信息量的概念，给出了必要属性的属性特征刻画，以及信息量与属性约简的关系。进一步基于信息量给出了集值信息系统的属性重要性，提出了集值信息系统属性约简的算法。通过实例验证了该算法的有效性。

参考文献：

[1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11: 341-356

[2] Pawlak Z, Grzymala-Busse J W, Slowinski R, *et al*. Rough sets[J]. Communication of the ACM, 1995, 38(11): 89-95

[3] Pawlak Z. Rough set theory and its application to data analysis[J]. Cybernetics and Systems, 1998, 9: 661-668

[4] 张文修，吴伟志，梁吉业等. 粗糙集理论与方法[M]. 北京：科学出版社，2001
Zhang W X, Wu W Z, Liang J Y, *et al*. Rough Set Theory and Approaches[M]. Beijing: Science Press, 2001

[5] 张文修，梁怡，吴伟志. 信息系统与知识发现[M]. 北京：科学出版社，2003
Zhang W X, Liang Y, Wu W Z. Information System and Knowledge Discovery[M]. Beijing: Science Press, 2003

[6] Wang S K M, Ziarko W. On optimal decision rules in decision tables[J]. Bulletin of Polish Academy of Sciences, 1985, 33: 693-696

[7] Miao D Q, Wang J. Information-based algorithm for reduction of knowledge[C]// IEEE International Conference on Intelligent Processing Systems, 1997: 1155-1158

[8] 苗夺谦，王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报，1999, 10(2): 113-116
Miao D Q, Wang J. An information representation of the concepts and operations in rough set theory[J]. Journal of Software, 1999, 10(2): 113-116

[9] 苗夺谦，胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展，1999, 36(6): 681-684
Miao D Q, Hu G R. A heuristical algorithm for reduction of knowledge[J]. Journal of Computer Research & Development, 1999, 36(6): 681-684

[10] Guan J W, Bell D A, Guan Z. Matrix computation for information systems[J]. Information Sciences, 2001, 131: 129-156

Attribute Reductions in Set-valued Information Systems

MA Jian-min[1], ZHU Chao-hui[2]

(1- Department of Mathematics and Information Sciences, Faculty of Science, Chang'an University, Xi'an 710064; 2- Research Institute of Shenzhen ZhuoCheng Concrete Module, Shenzhen 518000)

Abstract: Attribute reduction is one of important topics in the rough set theory. This paper mainly studies attribute reduction in set-valued information systems. Firstly, the information quality based on a preorder relation is defined in a set-valued information system. The judgment approach of attribute characterizations and relationships between the information quality and attribute reduction is discussed. On the basis of information quality, the significance of attributes is then introduced. And the relationship between the significance of attributes and attribute reduction is also investigated. Based on the information quality and significance of attributes, a heuristic algorithm for obtaining attribute reductions is presented, and the time complexity of the algorithm is then analyzed. By an example, we show this algorithm is effective.

Keywords: set-valued information system; preorder relation; information quality; significance of attribute; attribute reduction

Received: 20 July 2009. **Accepted:** 27 Feb 2010.

Foundation item: The National Natural Science Foundation of China (10901025); the Special Fund for Basic Scientific Research of Central Colleges (CHD2009JC028).